# **Supporting Information:** Coupling of diversification and pH adaptation during the evolution of terrestrial Thaumarchaeota

## SI Materials and Methods

### *Preparation of amoA alignments merging 454-derived and global diversity*

To test if the 454-sequences we used for phylogenetic analyses represented the published thaumarchaeotal *amoA* diversity, a global database was constructed by retrieving all available thaumarchaeotal *amoA* sequences in the National Center for Biotechnology Information (NCBI). The database was constructed by using BLASTn (1) employing reference *amoA* sequences from *Nitrosopumilus maritimus* (2), *Nitrosotalea devanaterra* (3) and *Nitrososphaera viennensis* (4) as representatives of the AOA 1.1a, 1.1a-associated and 1.1b lineages respectively. Sequences affiliated to *Nitrosocaldus yellowstonii* (5) were not considered as they fall outside the 1.1b, 1.1a and 1.1a-associated Thaumarchaeota (5). The thaumarchaeotal sequences were merged and aligned using MAFFT (6). Sequences that contained degenerate bases and/or were shorter than 400-bp were removed from the alignment. After trimming the alignment to the length of the 370-sequence 454 *amoA* alignment (see main materials and methods), sequences containing stop codons were removed. Finally, this alignment was dereplicated at an 89% cut-off using Uclust (7) resulting in an alignment of 270 unique sequences. A deeper-cut off was required using the global *amoA* data compared to the 454 sequences (where a 98% Uclust cut-off was employed), owing to the larger number of sequences retained, which made it impossible to run phylogenetic analyses in BEAST in a reasonable time. Therefore, to avoid under-estimation of modern diversity, the final 454 dataset (370 sequences) (8) was merged with the global data (270 sequences) and the non-redundant sequences were retained. We performed recombination analyses using RDP4 (9) and this led to an alignment of 613 *amoA* sequences, broadly representing global *amoA* diversity, which (along with a separate 370-sequence 454 alignment), was used in phylogenetic analyses, as described in the main materials and methods.

### *Preparation of 16S rRNA alignments merging 454-derived and global diversity*

Thaumarchaeotal sequences were downloaded from the Silva database (10) (SSURefNR99_119) based on their affiliation to one of the three thaumarchaeotal groups of interest (Marine group-I=MGI, South African Gold Mine Group 1=SAGMCG-1 and Soil Crenarchaeotic Group =SCG) within the published phylogenetic tree. Despite the high quality of this database, a further cleaning step was performed by deleting sequences with degenerate bases. These global sequences were then combined with the 16S rRNA data from our 454-sequencing (11), sequences were aligned using MAFFT (6) and the alignment was trimmed to the length of our 454 sequence data (11). We then performed recombination analyses using RDP4 (9), confirmed an absence of mutational saturation in the alignment (12) and dereplicated the sequences at a 97% cut-off using Uclust (7). In order to perform the most direct comparison to our *amoA* analyses, we aimed to include Thaumarchaeota possessing *amoA* genes (1.1b, 1.1a and 1.1a-associated lineages), although we acknowledge that the presence of an *amoA* gene does not necessarily imply an energy metabolism driven by ammonia oxidation. We selected relevant 16S rRNA genes by performing a neighbour-joining phylogenetic analysis in Mega 6 (13) using a maximum composite likelihood distance model (14) assuming a gamma distribution of among-site rate distribution. This analysis incorporated reference 16S rRNA sequences that fall outside 1.1b, 1.1a and 1.1a-associated lineages 3 (15). Subsequently, any sequences that grouped within these clusters were removed, leading to a final alignment of 508 sequences that was used in recombination and phylogenetic analyses, as described in the main materials and methods.

## SI text

### *Recombination analysis*

Using the software RDP4 (9) with stringent criteria, we identified 55 *amoA* sequences affected by recombination in the dereplicated alignment of 425 sequences (Table S1). By matching these sequences to the original set of 108,192 sequences, we estimate that 8,655 (8.02%) were affected by

recombination. However, 99.5% of these events occurred within the alkalinophilic *Nitrososphaera* C12 cluster (representing 99.98% of the sequences within this cluster). These 55 sequences were removed from the main alignment used in all phylogenetic analyses.

**SI references**


1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

2. Könneke M *et al*. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* 437(7058):543–546.

3. Lehtovirta-Morley LE, Stoecker K, Vilcinskas A, Prosser JI, Nicol GW (2011) Cultivation of an obligate ammonia oxidizer from a nitrifying acid soil. *Proc Natl Acad Sci USA* 108(38):15892-15897.

4. Tourna M e*t al* (2011) *Nitrososphaera viennensis*, an ammonia oxidizing archaeon from soil. *Proc Natl Acad Sci USA* 108(20):8420-8425.

5. de la Torre JR, Walker CB, Ingalls AE, Könneke M, Stahl DA (2008) Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. *Environ Microbiol* 10(3):810-818.

6. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772-780.

7. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461.

8. Gubry-Rangin C *et al*. (2011) Niche specialization of terrestrial archaeal ammonia oxidizers. *Proc Natl Acad Sci USA* 108(52):21206-21211.

9. Martin DP *et al*. (2010) RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26(19): 2462–2463.

10. Yilmaz P *et al*. (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucl Acids Res* 42:D643-D648.

11. Vico Oton E, Quince C, Nicol GW, Prosser JI, Gubry-Rangin C (2015) Phylogenetic congruence and ecological coherence in terrestrial Thaumarchaeota. *ISME J* (in press).

12. Xia X, Xie Z, Salemi M, Chen L, Wang Y (2003) An index of substitution saturation and its application. *Mol Phylogenet Evol* 26(1):1-7.

13. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol Biol Evol* 30(12):2725-2729.

14. Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci USA* 101(30):11030-11035.

15. Nicol GW, Tscherko D, Embley TM, Prosser JI (2005) Primary succession of soil Crenarchaeota across a receding glacier foreland. *Environ Microbiol* 7(3):337-347.


**SI figures**

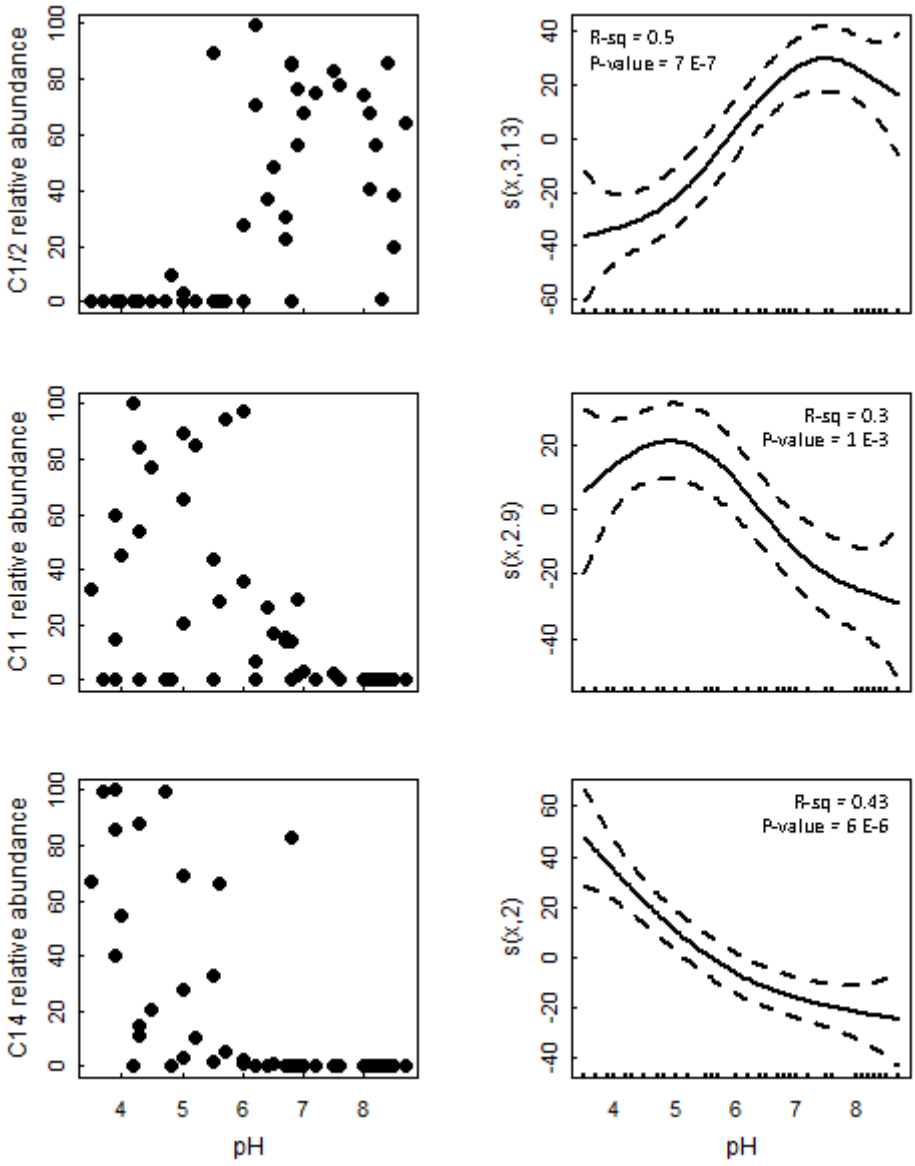


**Fig. S1.** The relative abundance of the sequences within each of the three abundant clusters is represented as a function of soil pH. The first column represents the percentage relative abundance and the second column the best-fitting model of this distribution according to generalized additive modelling. The regression coefficients and associated *p*-values are provided.

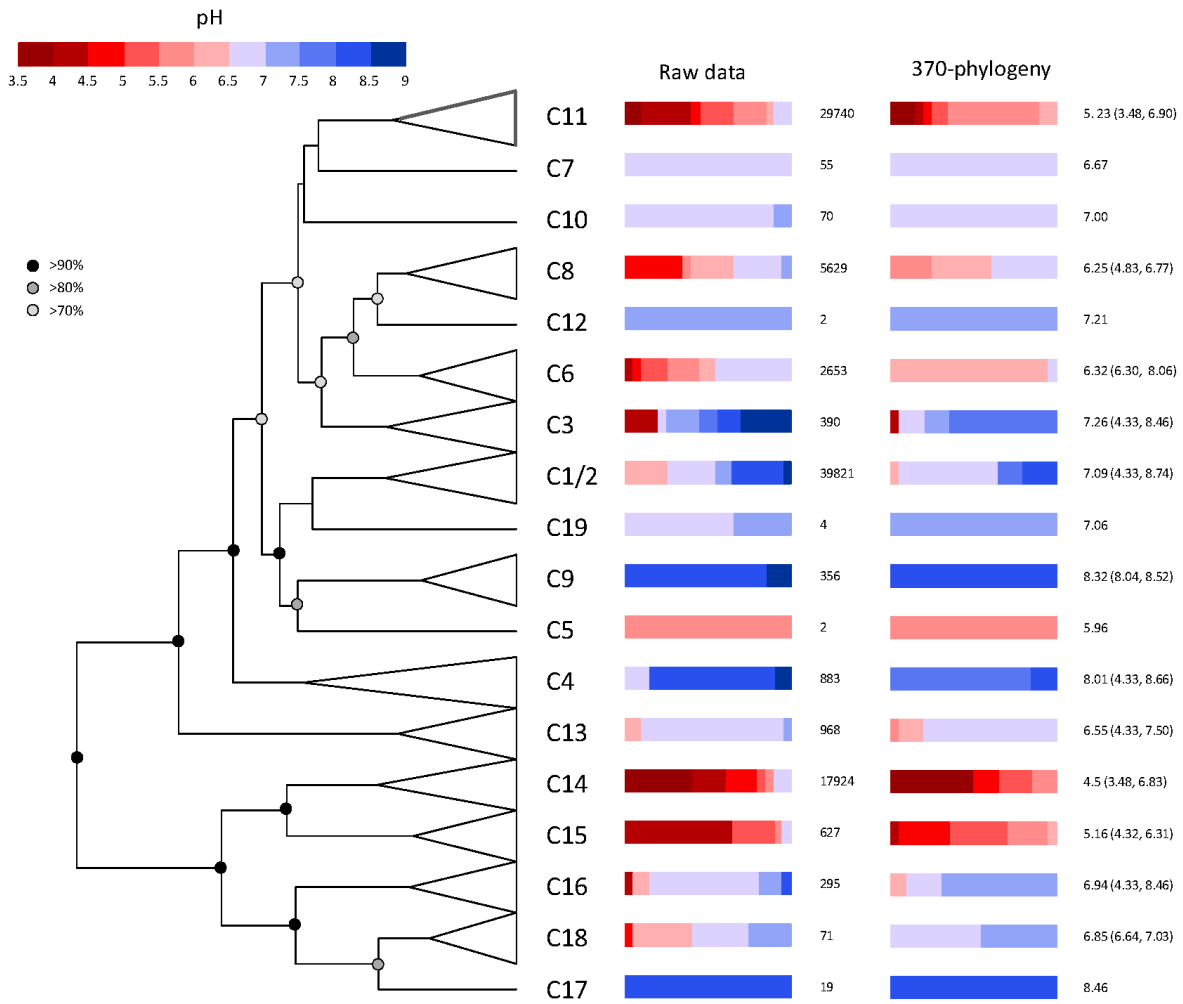


**Fig. S2.** *amoA* cladogram highlighting 18 defined phylogenetic clusters with associated bar plots showing the distribution of sequences according to soil pH. The first column of bar plots highlights the sequence distribution based on the number of un-dereplicated 454 reads within each cluster (numbers at the right of the bar indicate the total number of sequences). The second column of the bar plots highlights the mean pH calculated for each tip of the final phylogenetic tree used in all analyses. In this case, the numbers at the right of the bar represent mean pH values with lower and upper means observed for the tips within the cluster. Both approaches indicate similar pH preference for each cluster.

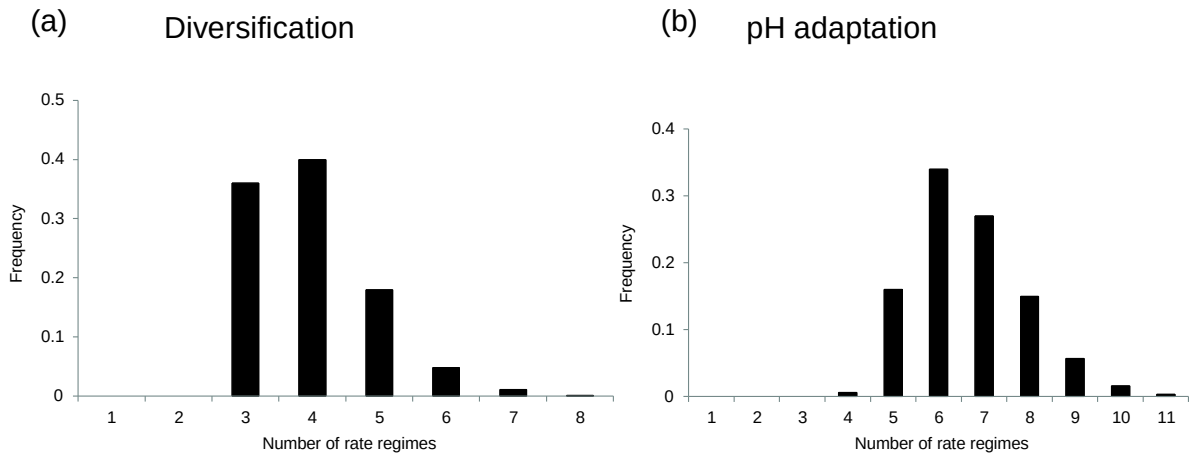


**Fig. S3.** Marginal distributions of the number of distinct evolutionary rate regimes for (a) rates of diversification and (b) rates of pH adaptation based on a sample of 5000 phylogenies taken from the BAMM posterior distribution.

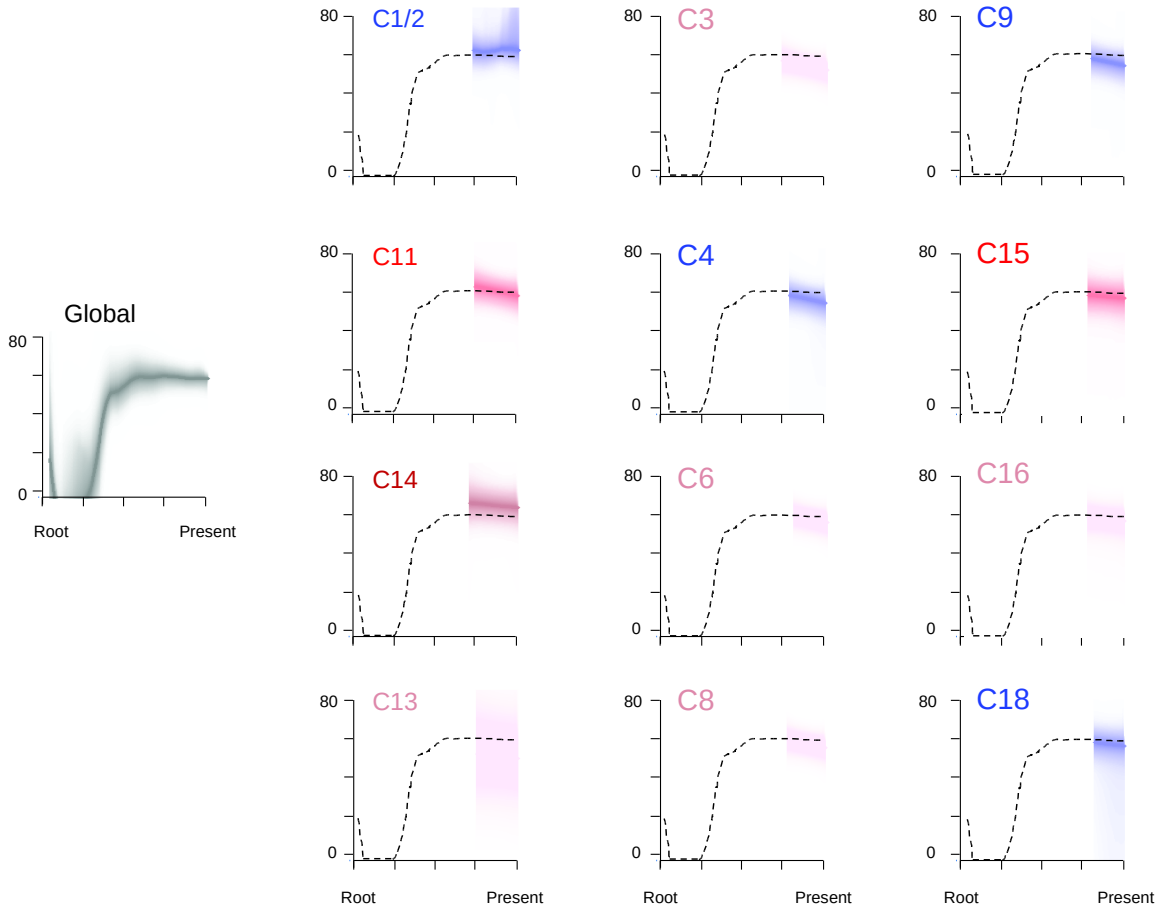


**Fig. S4.** BAMM plots of net-diversification for several specific thaumarchaeotal clusters, as presented in Fig. 2. For each cluster, the global rate is plotted for comparison (black dotted line). The color for each cluster is linked with its pH specialization.

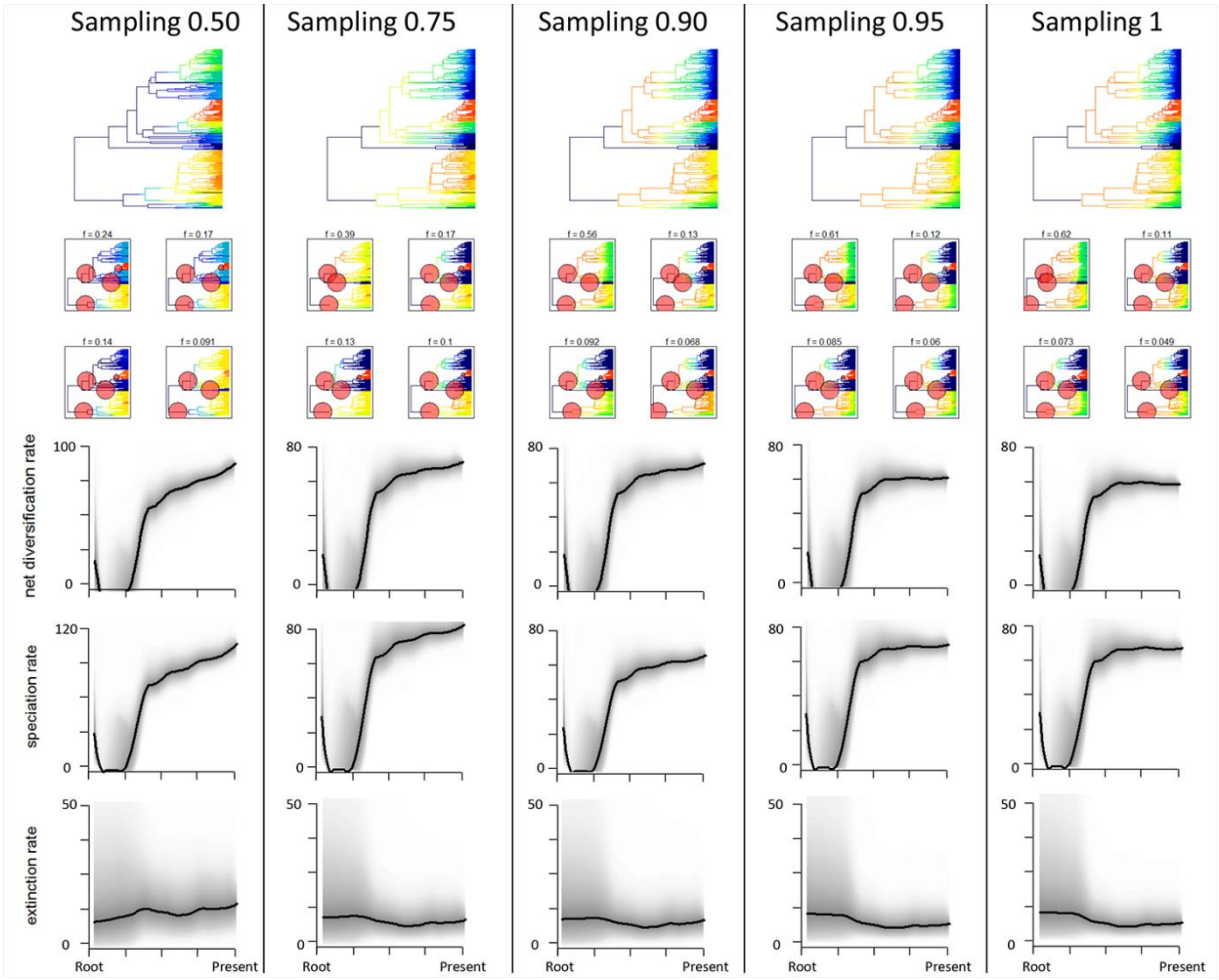


**Fig. S5.** Effect of random incomplete sampling of global lineage diversity on the BAMM diversification analysis. The BAMM analysis was performed assuming that the overall diversity in our samples represented a range of different proportions (50, 75, 90, 95 or 100%) of the true global diversity. This effect was tested on (a) the heterogeneity in diversification rates along each branch of the thaumarchaeotal phylogeny, (b) the four most probable diversification rate-shift configurations along with their individual contributions to the posterior distribution of all sampled BAMM models, (c, d and e) the respective rates of net diversification, speciation and extinction through time across the radiation of terrestrial Thaumarchaeota. Other details are as described in the Fig. 2 legend.

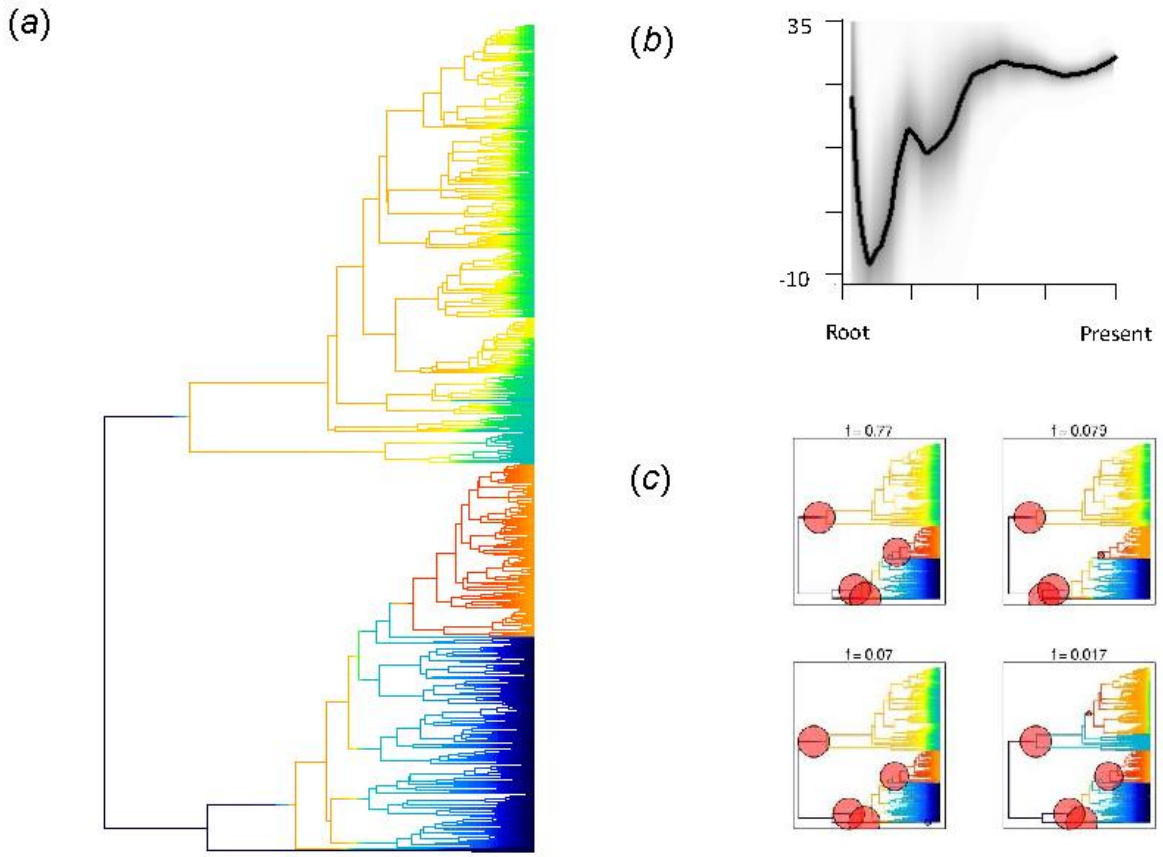


**Fig. S6.** Evolutionary heterogeneity in diversification rates in terrestrial Thaumarchaeota inferred from the global *amoA* tree within the BAMM framework, based on an alignment including the available published *amoA* and the present 454 sequence diversity. All details are as described in the Fig. 2 legend, except that only the plot of net diversification rate is represented in *(b)*.

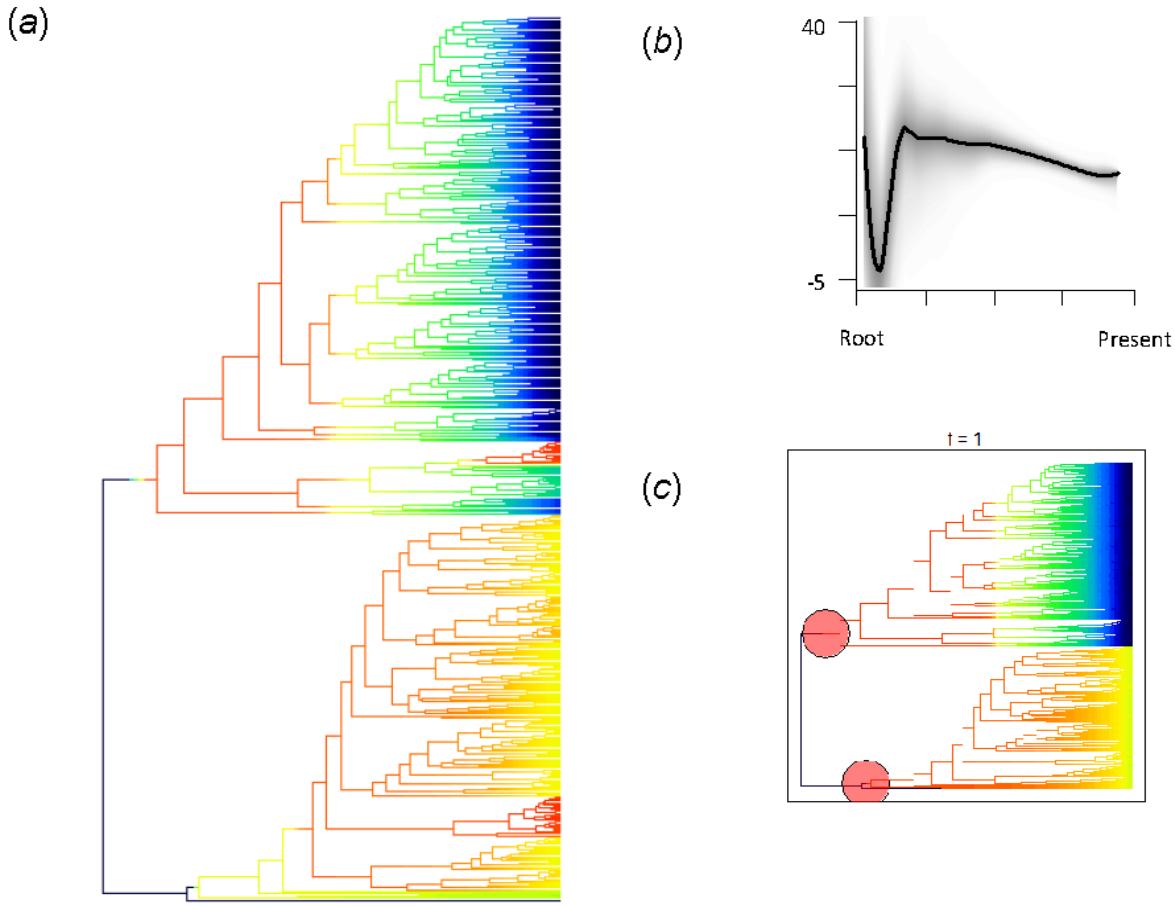


**Fig. S7.** Evolutionary heterogeneity in diversification rates in terrestrial Thaumarchaeota inferred from the global 16S rRNA tree within the BAMM framework, based on an alignment including the available published 16S rRNA diversity merged with 454 sequences from another study (11). All details are as described in the Fig. 2 legend, except that only the plot of net diversification rate is represented in *(b)*.

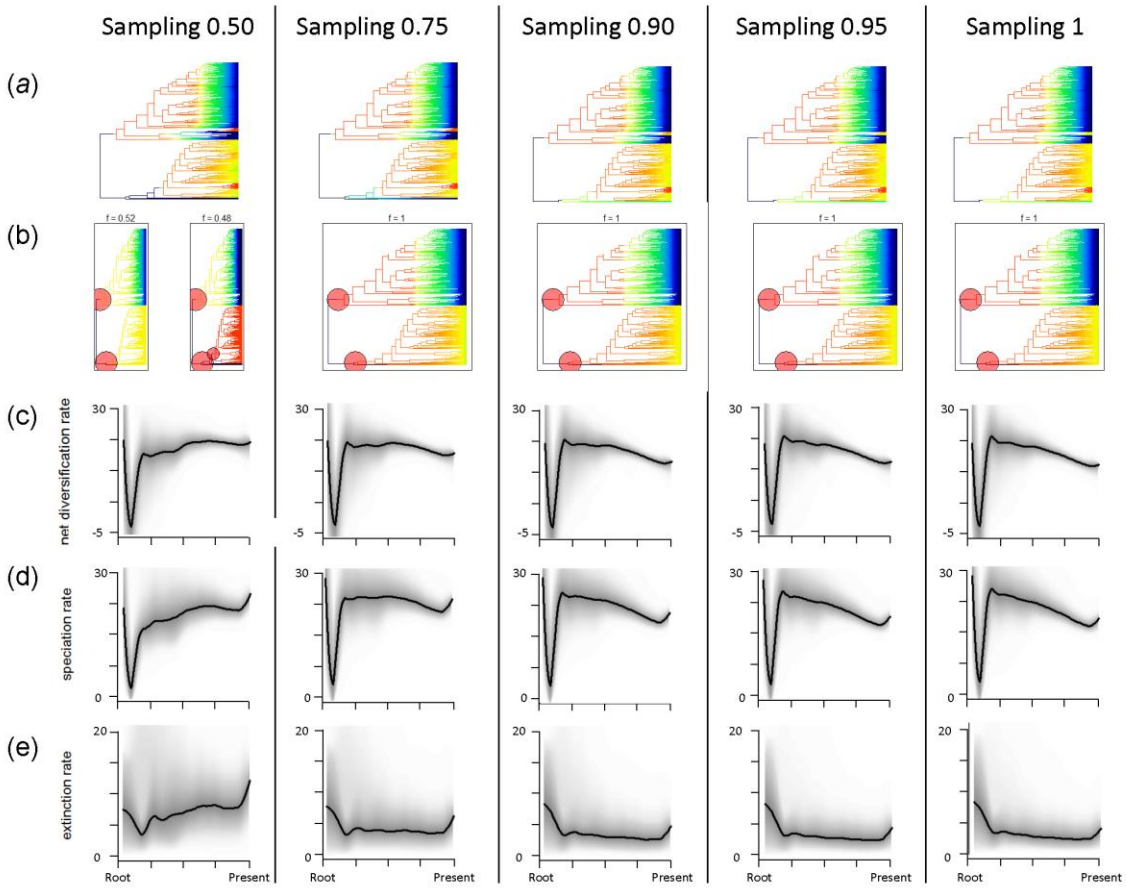


**Fig. S8.** Effect of random incomplete sampling of global lineage diversity on the BAMM diversification analysis for the global 16S rRNA tree. All details are as described in the Fig. S5 legend.

**SI table**

**Table S1.** Recombinant sequences detected by the RDP analysis ($p$<0.01 cut-off) after manual curation. Only events detected by more than 3 methods were considered (see methods). [D] indicates that two recombination events were detected in the recombinant sequence whereas [T] implies that the recombinant sequence only contained trace evidence of recombination (p>0.01). The * subscript indicates that the breakpoint position is undetermined.

| Event number | Recombinant Sequence(s) | Recombinant Sequence(s) with similar event detected | Cluster affiliation of parental sequences | | Breakpoint Positions | | Detection Methods | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | *Major* | *Minor* | *Begin* | *End* | *RDP* | *GENECONV* | *Bootscan* | *Maxchi* |
| 1 | C15_S17_281_2 | | C14 , C15 | C11 | 1* | 82 | 3.59E-16 | 1.44E-14 | 3.63E-16 | 1.28E-04 |
| 2 | C11_S40_141_2 | | C4, C6, C8, C10, C11 | C14, C15 | 492 | 582* | 1.37E-14 | 3.57E-14 | 7.81E-16 | 6.43E-05 |
| 3 | C12_S27_352_3 D | | C12 | C1 | 374 | 582* | 1.13E-12 | 1.60E-10 | 8.53E-13 | 4.32E-08 |
| 4 | C8_S45_463_2 D | | C8 | C2 | 328 | 582* | 2.77E-12 | 4.36E-09 | 1.98E-12 | 8.25E-09 |
| 5 | C7_S10_195_2 | | C7 | C2 | 1* | 184 | 1.98E-12 | 2.63E-10 | 6.21E-13 | 1.21E-07 |
| 6 | C11_S41_521_2 | | C11 | C13 | 1* | 116 | 8.86E-10 | 2.26E-08 | 1.09E-08 | NS |
| 7 | C12_S11_111_2 | 31 other C12 sequences [(a)] | C12 | C1 | 1* | 196 | 1.91E-10 | 1.71E-08 | 7.78E-10 | 5.26E-06 |
| 8 | C9_S27_300_3 | | C9 | C2 | 458 | 582* | NS | 1.35E-09 | 2.83E-11 | 3.35E-04 |
| 9 | C8_S23_546_2 | C8_S23_141_2 | C8 | C2 | 1* | 183 | 1.93E-09 | 2.04E-08 | 3.24E-09 | 1.54E-05 |
| 10 | C1_S26_142_3 | C1_S14_251_2 | C1 | C12 | 1* | 196 | 2.47E-09 | 9.03E-08 | 2.48E-09 | 1.29E-06 |
| 11 | C2_S23_96_3 | C2_S44_47_3T | C2 | C6, C11 | 504 | 582* | 2.90E-09 | 6.11E-08 | 1.24E-09 | 1.97E-03 |
| 12 | C11_S42_162_2 | C11_S41_495_2 | C11 | C6 | 1* | 177 | 3.22E-09 | 1.30E-06 | 5.21E-09 | 2.15E-05 |
| 13 | C1_S5_67_2 | | C1 | C2 | 402 | 582* | 3.31E-10 | 5.53E-08 | 1.09E-10 | 4.93E-07 |
| 14 | C2_S46_107_2 | | C2 | C13 | 1* | 219 | 2.19E-11 | 4.34E-10 | 3.74E-10 | 7.79E-07 |
| 15 | C8_S10_301_2 | C8_S45_463_2 D | C8 | C2 | 450 | 582* | 2.02E-09 | 4.29E-08 | 4.27E-10 | 2.55E-05 |
| 16 | C4_S37_135_2 | | C4 | C9 | 1* | 124 | 1.39E-07 | 1.53E-06 | 4.23E-08 | 6.23E-03 |
| 17 | C2_S27_646_2 | | C2 | C1 | 396 | 582* | 3.30E-07 | 3.07E-04 | 6.34E-07 | 3.34E-06 |
| 18 | C2_S20_243_2 | | C2 | C8, C11 | 488 | 582* | 3.81E-07 | 2.54E-06 | 5.65E-08 | 4.96E-03 |
| 19 | C15_S1_23_3 | | C14 | C14 | 1* | 192 | 2.29E-06 | 5.11E-05 | 7.73E-07 | 1.20E-03 |
| 20 | C12_S11_102_3 D | | C12 | C4 | 508 | 582* | NS | 2.43E-05 | 2.47E-05 | 2.32E-03 |
| 21 | C6_S10_273_2 | | C6 | C11 | 478 | 582* | 1.63E-04 | 7.82E-04 | 9.00E-04 | NS |
| 22 | C11_S9_238_2 | | C11 | C11 | 400 | 582* | 7.33E-04 | 5.05E-02 | 2.69E-04 | 2.08E-04 |

(a) C12_S11_0_93, C12_S11_102_3 [D], C12_S11_122_2, C12_S11_140_4, C12_S13_101_2, C12_S13_59_3, C12_S13_95_2 [T], C12_S21_275_2, C12_S21_565_2, C12_S27_352_3 [T] [D], C12_S27_610_3, C12_S27_809_2, C12_S31_198_2, C12_S37_11_10, C12_S37_155_2, C12_S37_157_2, C12_S37_184_2, C12_S37_199_2, C12_S37_210_2, C12_S37_221_3, C12_S37_286_2, C12_S37_303_2, C12_S37_320_2, C12_S37_340_2, C12_S37_406_2, C12_S37_408_2, C12_S37_453_2, C12_S37_461_2, C12_S37_476_2, C12_S37_478_2, C12_S7_241_3[T]

**Table S2.** Contextual data for the soils used in this study. The site number relates to a previous study (9) and the different parameters measured are: pH, percentage of organic matter (loss on ignition (LOI)), phosphorous concentration (Olsen extractable $PO_4$ in mg $kg^{-1}$) and the concentration (in mg $kg^{-1}$) of six metals (Al, Mo, Mn, Hg, Cu and Zn).

| Site | pH | Aluminium (mg/kg) | Organic matter (%) | Phosphorus (mg/kg) | Molybdenum (mg/kg) | Manganese (mg/kg) | Mercury (mg/kg) | Copper (mg/kg) | Zinc (mg/kg) |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 4.83 | 8730 | 4.88 | 6.9 | 0.88 | 878 | 0.03 | 14.50 | 44.8 |
| 6 | 8.25 | 4720 | 24.10 | 17.6 | 0.33 | 723 | 0.10 | 8.47 | 80.0 |
| 9 | 4.16 | 8110 | 5.75 | 3.0 | 0.55 | 462 | 0.03 | 8.62 | 84.0 |
| 11 | 8.06 | 12000 | 5.82 | 20.6 | 0.42 | 205 | 0.03 | 16.90 | 65.2 |
| 12 | 5.65 | 15100 | 24.00 | 25.0 | 1.85 | 173 | 0.12 | 31.10 | 61.7 |
| 13 | 8.40 | 8520 | 6.37 | 10.3 | 0.56 | 504 | 0.03 | 8.84 | 56.0 |
| 16 | 4.33 | 4530 | 85.08 | 14.0 | 1.21 | 32 | 0.25 | 16.40 | 31.2 |
| 18 | 3.70 | 883 | 15.01 | 9.0 | 0.46 | 18 | 0.03 | 6.57 | 20.7 |
| 21 | 8.16 | 10000 | 3.72 | 20.0 | 0.56 | 343 | 0.03 | 11.20 | 53.1 |
| 22 | 3.93 | 1990 | 19.89 | 6.9 | 2.85 | 11 | 0.14 | 24.10 | 13.4 |
| 23 | 6.18 | 13400 | 5.78 | 10.6 | 1.31 | 782 | 0.03 | 31.00 | 103.0 |
| 26 | 8.05 | 12500 | 5.10 | 40.1 | 1.29 | 1360 | 0.03 | 27.90 | 105.0 |
| 27 | 8.04 | 11100 | 5.06 | 4.1 | 0.89 | 2910 | 0.03 | 91.20 | 174.0 |
| 29 | 3.85 | 5320 | 59.04 | 33.0 | 0.77 | 92 | 0.16 | 14.60 | 52.3 |
| 30 | 6.77 | 13600 | 4.76 | 39.2 | 0.34 | 610 | 0.03 | 16.50 | 79.8 |
| 31 | 7.21 | 8140 | 4.51 | 68.2 | 0.65 | 612 | 0.23 | 34.40 | 170.0 |
| 34 | 5.96 | 10600 | 11.78 | 3.4 | 4.77 | 654 | 0.09 | 12.10 | 61.5 |
| 35 | 6.86 | 3560 | 3.20 | 68.0 | 0.20 | 86 | 0.03 | 3.84 | 15.7 |